MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1533
C.B.C.L. Paper No. 120

MARCH, 1995

Spatial Reference Frames for Object Recognition

Tuning for Rotations in Depth

N.K. Logothetis, J. Pauls, and T. Poggio

nikos@bcmvision.bcm.tmc.edu, jpauls@bcmvision.bcm.tmc.edu, tp-temp@ai.mit.edu

Abstract

The inferior temporal cortex (IT) of monkeys is thought to play an essential role in visual object recognition. Inferotemporal neurons are known to respond to complex visual stimuli, including patterns like faces, hands, or other body parts. What is the role of such neurons in object recognition? The present study examines this question in combined psychophysical and electrophysiological experiments, in which monkeys learned to classify and recognize novel visual $3D$ objects. A population of neurons in IT were found to respond selectively to such objects that the monkeys had recently learned to recognize. A large majority of these cells discharged maximally for one view of the object, while their response fell off gradually as the object was rotated away from the neuron's preferred view. Most neurons exhibited orientation-dependent responses also during view-plane rotations. Some neurons were found tuned around two views of the same object, while a very small number of cells responded in a view-invariant manner. For five different objects that were extensively used during the training of the animals, and for which behavioral performance became view-independent, multiple cells were found that were tuned around different views of the same object. No selective responses were ever encountered for views that the animal systematically failed to recognize. The results of our experiments suggest that neurons in this area can develop a complex receptive field organization as a consequence of extensive training in the discrimination and recognition of objects. Simple geometric features did not appear to account for the neurons' selective responses. These findings support the idea that a population of neurons – each tuned to a different object aspect, and each showing a certain degree of invariance to image transformations – may, as an assembly, encode complex $3D$ objects. In such a system, several neurons may be active for any given vantage point, with a single unit acting like a blurred template for a limited neighborhood of a single view.

This paper describes research done at the Baylor College of Medicine. and the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. **Nikos K. Logothetis** was supported by the contract N000 14-93-1-0209 of the Office of Naval Research (1992) and the McKnight Endowment Fund for Neuroscience (1993). **Tomaso Poggio** was supported by the Office of Naval Research contract N00014-93-1-0385, and by the NSF grant ASC-92-17041.

1 Introduction

Object recognition can be thought of as the process of matching the image of an object to its representation stored in memory. Because different viewing, illumination, and context conditions generate different retinal images, the nature of the stored representation and the process of normalization of the sensory input presents one of the greatest challenges to understanding biological recognition. It is well known that familiar objects are recognized regardless of viewing angle, scale or position in the visual field. How is such perceptual object constancy accomplished? Does the brain transform the sensory or the stored representation to discard the image variability resulting from different viewing conditions, or does generalization occur as a consequence of perceptual learning, that is, of being acquainted with different instances of any given object? The present paper addresses one aspect of this issue, namely, how the primate recognition system may compensate for changes in viewing angle and distance, ignoring the image changes resulting from variation of the illumination and context. Moreover, the issue is addressed at the level of subordinate categorizations of objects.

Studies indicate that objects can be identified at a number of levels of abstraction, but are most easily recognized at what is referred to as the *basic level* (Rosch et al., 1976). For instance, a *barn swallow* is perceived first as a *bird* , rather than as a *swallow* or an *Avian.* Classifications above the basic level are more general and are called *superordinate.* In contrast, *subordinate* level refers to classifications below the basic level and are more specific, sharing a great number of attributes with other members of the object class. The behavioral performance of humans for subordinate classifications is strongly view dependent (Rock and DiVita, 1987; Tarr and Pinker, 1990; Edelman and Bülthoff, 1992), presumably because it largely relies on the recognition of subtle differences in the shape of complex objects. It is also this type of classification that is most seriously impaired by circumscribed damage to the human cerebral cortex (Damasio, 1990). It appears that, at least in humans, distinct shape differences may be the basis for reliable object recognition under any viewing conditions. Objects with distinct shape are easiest and fastest recognized whether of a basic-level or not. For instance a *penguin*, i.e. an atypical exemplar the basic-level category *birds*, is most likely to be first recognized as "penguin" rather than as a "bird", a classification termed *entry level* recognition (Jolicoeur et al., 1984). Penguins do indeed have a distinct shape when compared with most other animals, but also differ a great deal from any other bird.

Conceptual hierarchies like those mentioned above reflect certain types of interactions between the human perceiver and objects in the environment. As such they also reflect the "default" probabilities of the required discriminations for any given class of objects. Thus in a domain of expertise, subordinate-level categories may be as differentiated as the basic-level categories, and the former categorizations may be as fast as the latter (Tanaka and Taylor, 1991). Clearly, in the nonhuman primate categories have no bearing on language. Nonetheless, there is little doubt that monkeys are capable of categorizations of objects like *predators, prey, infant monkeys*, or *food*: categories of objects usually having distinct shape differences. It has also been shown that monkeys can be trained to be "experts" in discriminations of objects of a novel class, the members of which share great shape similarities (Logothetis et al., 1994). It is this latter type of object discriminations that was used to study the spatial reference system of object representations in the non-human primate and the activity of neurons in the temporal cortex during the execution of the recognition task.

The reference system used in matching object shapes to their representations encoded in visual memory is a key question in the research of visual object recognition (Farah, 1985; Ullman, 1989; Tarr and Pinker, 1989). Theories relying on object-centered representations assume either a complete three-dimensional description of an object (Ullman, 1989), or a structural description of the image that specifies the relationships among viewpoint-invariant volumetric primitives (Marr, 1982; Biederman, 1987). Whereas such theories correctly predict the view-independent recognition of familiar objects (Biederman, 1987), they fail to account for performance in recognition tasks with of novel objects at the subordinate level (Rock & DiVita, 1987; Rock et al., 1981; Tarr & Pinker, 1990; Bülthoff and Edelman, 1992; Edelman & Bülthoff, 1992). Viewpoint-dependent, image-based models, on the other hand, represent three-dimensional objects as a set of $2D$ views, or aspects, and recognition consists of matching image features against the views in this set.

Although such models can account for the performance of human subjects in any recognition task, they are usually considered implausible because of the memory a system would require to store all discriminable views of many objects. These objections, however, have recently been challenged by computer simulations showing that a simple network can recognize $3D$ objects by *interpolating* between a small number of stored views (Poggio and Edelman, 1990; Logothetis et al., 1994). This network (Figure 1) uses a small set of sparse data, corresponding to an object's training views, to synthesize an approximation of a multivariate function (Poggio and Girosi, 1990) representing the object.

In such a network a view can be represented by a set of any image features, such as the orientations or positions of object parts, shape metrics, texture, or color. Complex features can be created hierarchically from simpler ones as shown in Figure 1. The performance of the network was tested with geometrical features like the position of the vertices of wire-objects (Poggio & Edelman, 1990), or their orientations (Logothetis et al., 1994), or with features extracted from real images of wire-objects (Brunelli and Poggio, 1991b) or faces (Brunelli and Poggio, 1991a). The actual features used by a biological recognition system are presently unknown and their nature is an important experimental question *per se.* Nonetheless, some of the arbitrary features used in the simulations can provide a measure of object similarity.

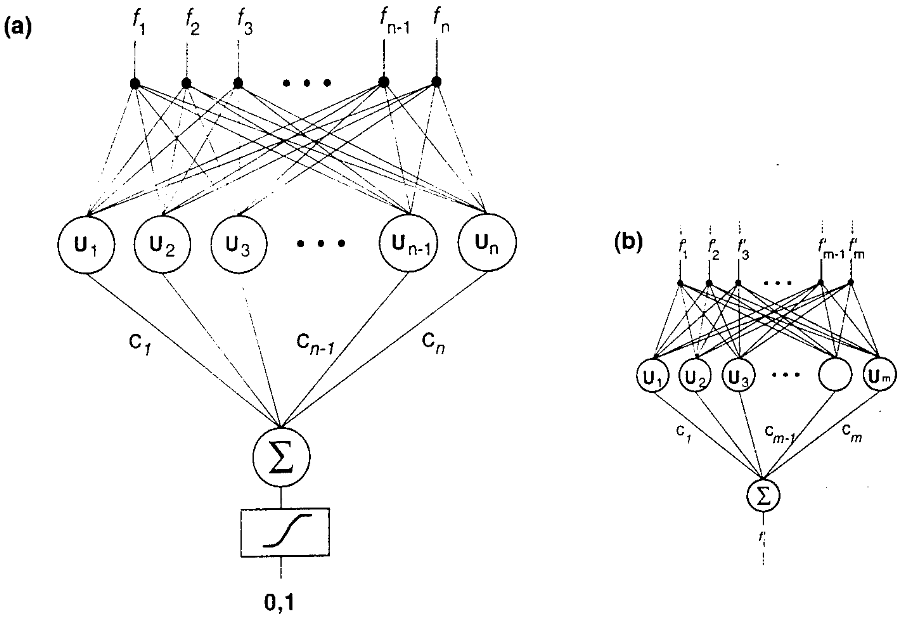

Figure 1: (a) Performance of a regularization network trained with the $0^o, 60^o, 120^o$, and 180^o views of an wire-object. Each "hidden-layer" unit takes a similarity-measure between a novel view and a template stored in the unit's memory, by calculating the euclidean distance $||\mathbf{V} - \mathbf{T}_i||$ of the input vector $\mathbf{V}$ from its learned view $\mathbf{T}_i$, and subsequently computing the function $F_i(\mathbf{V}) = exp(-||\mathbf{V} - \mathbf{T}_i||^2)$ of this distance. The activity of the entire network is conceived of as the weighted, sum of each unit's output ($F(\mathbf{V}) = \sum_{i=1}^{N} c_i exp(-||\mathbf{V} - \mathbf{T}_i||^2)$). A decision criterion can be applied for yes/no type of performance. The basic scheme can be hierarchically used for composing complex features out of simpler ones (small inset).

Based on such features, simple simulations argue against the implausibility of a view-based recognition system.

Also in agreement with the basic idea that a limited number of views might be sufficient to accomplish view-invaraince, are recent psychophysical experiments showing that human subordinate-level recognition performance can be best predicted by assuming that subjects *interpolate* between familiar object views (Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992). Similar performance has been observed in nonhuman primates performing a subordinate level recognition task (Logothetis et al., 1994). It was shown that monkey's were limited in their ability generalize recognition to novel views of an object, performing best for a most familiar view and gradually worse for views with increasing distance from the known view. Familiarity with two views of an object allowed the interpolation of recognition between the views if they were close enough together, say 75^o apart, but resulted in two independent regions of generalization if they were far apart, say 160^o. In most cases, however, only three to five familiar views were needed for the animal to achieve view-invariant performance around one axis.

A recognition architecture that could underlie such performance might rely on small-scale networks with units that are broadly tuned to views or features of a learned object. Neurons responding to complex $2D$ patterns, including face or hand views (Gross et al., 1972; Bruce et al., 1981; Rolls, 1984; Desimone et al., 1984; Yamane et al., 1988), have indeed been reported in inferotemporal cortex of the monkey by different researchers (Richmond et al., 1987; Miyashita, 1988: Tanaka et al., 1991; Fujita et al.. 1992). Such cells discharge more strongly to complex patterns than to any simple stimulus, and are found even in the earliest stages of ontogeny of the primate (Rodman et al., 1993). A detailed investigation of the cells showing high selectivity for faces has revealed several different types or classes of neurons in the superior temporal sulcus, each broadly tuned to one view of the head, *e.g.* full face or profile (Perrett, 1985). Similarly, neurons have been reported that respond selectively to static or dynamic information about the body, or body parts, some of which were dependent on the observer's vantage point (Perrett et al., 1989; Wachsmuth et al., 1994). Is such a configurational selectivity specific only for faces or body parts, or can it be generated for any novel object as a result of extensive training?

Clinical observations have shown that the recognition of living things can be selectively impaired (Farah et al., 1991). This may imply that the perception of faces or biological forms in general is mediated by specialized neural populations. If so, then the complex-pattern selectivity (faces, body parts, etc.) reported in the above studies may be unique to the representation of the class of "living things", with different encoding mechanisms responsible for the recognition of other objects. In general, objects may be represented by large populations of cells each encoding a simple feature, or the conjunction of simple features that are characteristic for a given class. Alternatively, a system based on neurons selective for complex configurations may provide one mechanism for encoding any object that cannot undergo much meaningful decomposition in the course of recognition. Some subordinate categorizations cannot rely on part decomposition. We are unlikely to recognize individual faces. for example. by simply detecting the existence of two eyes. the nose and the mouth, as each individual is likely to have the same parts in approximately the same positions. It is a holistic and/or a metric representation that probably underlies the recognition of a person's face. The same reasoning may apply for the recognition of individual objects of other classes, particularly artificial objects composed of similar parts. Thus. the question arises: If monkeys are extensively trained to identify novel $3D$ objects of a class whose members show a great deal of structural similarity, then would one find neurons in the brain which respond selectively to the views of such objects?

We have examined this possibility using two classes of novel. computer-rendered stimuli: Gouraud-shaded wire-like and amoeboid objects (Bülthoff & Edelman. 1992; Edelman & Bülthoff, 1992; Logothetis et al., 1994). The monkeys were trained in a matching task, generalized across translation. scaling and orientation changes. Within an object class the target-distractor similarity varied between one extreme, where distractors were generated by randomly selecting shape-parameters, such as the positions of vertices or protrusions, the sharpness of angles between segments, or the moment of inertia of the objects, and the other extreme. where distractors were generated by adding different degrees of noise to the parameters of the target. A variety of other digitized $2D$ or $3D$ patterns, *e.g.* . geometric objects, scenes, body-parts, were also used as controls in the physiological experiments.

2 Methods

2.1 Subjects and Surgical Procedures

Two juvenile rhesus monkeys (*Macaca mulatta*) weighing 7-9 kg were tested in the electrophysiological studies. The animals were cared for in accordance with the National Institutes of Health Guide, and the guidelines of the Animal Protocol Review Committee of the Baylor College of Medicine.

After preliminary training, the animal underwent a aseptic surgery. using isoflurane anesthesia (1.2% - 1.5%), for the placement of the head restraint post and the scleral search eye coil. Throughout the surgical procedure the heart rate, blood pressure and respiration were monitored constantly and recorded every 15 minutes. Body temperature was kept at 37 degrees using a heating pad. Postoperatively, the monkey was administered an opioid analgesic (Buprenorphine hydrochloride 0.02 mg/kg, IM) every 6 hours for one day, and Tylenol (10 mg/kg) and antibiotics (Tribrissen 30 mg/kg) for 3-5 days. At the end of the training period another sterile surgery was performed to implant a chamber for the electrophysiological recordings.

2.2 Animal Training

Standard operant conditioning techniques with positive reinforcement were used to train the monkey to perform the task. Initially, the animals were trained to recognize a target's zero view among a large set of distractors. When they had learned the zero view they were encouraged to generalize recognition to neighboring views resulting from progressively larger rotations around one axis. The criterion required before training with another object was 95% correct over a range of $\pm 90^o$ for the target. and less than 5% false alarm rate for all distractors. In the early stages of training several days were required to train the animals to perform the same task for a new object. Four months of training was required on average for the monkey to learn to generalize the task across different types of objects of one class, and about six months were required for the animal to generalize for different object classes.

The similarity of the targets to the distractors was gradually increased within an object class. In the final stage of the experiments distractor wire-objects were generated by adding different degrees of position or orientation noise to the target objects. A criterion of 95% correct for several objects was required to proceed with the psychophysical data collection.

In the initial training phase, the animal received continuous feedback about its performance. Each correct response was rewarded with a drop of juice. In the later stages of the training the animals were reinforced on a variable-ratio schedule which administered a reward after a specified average number of correct responses had been given. Finally, in the last stage of the behavioral training the monkey was rewarded only after ten consecutive correct responses. The end of the observation period was signalled with a full-screen. green light and a juice reward for the monkey. The variable-ratio schedule was also used throughout the period of psychophysical data collection.

During the behavioral training, independent of the reinforcement schedule, the monkey always received feedback as to the correctness of each response. Incorrect reports aborted the entire observation period. During psychophysical data collection, on the other hand, the monkey was presented with novel objects and no feedback was given during the testing period. The behavior of the animals was monitored continuously during the data collection by computing on-line hit rate and false alarms. Arbitrary performance or the development of hand-preferences, *e.g.* giving only right hand responses, was discouraged during psychophysical data collection by randomly interleaving sessions of actual data collection with sessions in which a novel object was presented but correct performance was required of the animal (i.e., incorrect responses resulted in aborts).

In the electrophysiological experiments the animal was required to maintain fixation throughout the entire observation period. Eye movements were measured using the scleral search coil technique and digitized at 200Hz.

2.3 Electrophysiological recording

Recording of single unit activity was done using Platinum-Iridium electrodes of 2-3 Megohms impedance. The electrodes were advanced into the brain through a guide tube mounted into a ball-and-socket positioner (Monkey S5396: AP = 15, L = 22; Monkey B63A

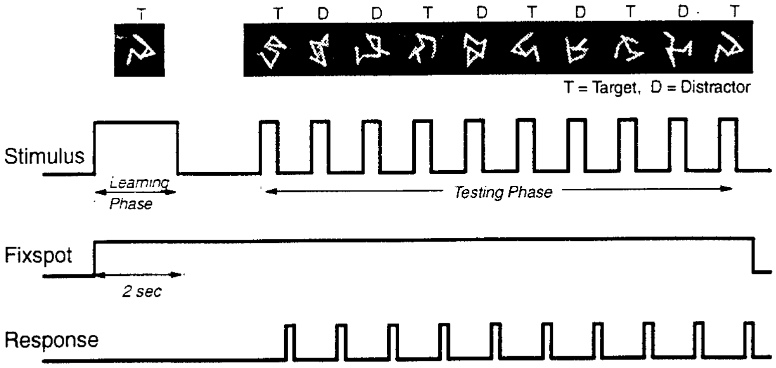

Figure 2: The experimental paradigm. Each observation period began with the presentation of a fixation spot. Successful fixation was followed by the learning phase, after which up to ten single, static views of either the target or a distractor were presented sequentially (testing phase). The subject was required to respond to each one in turn, indicating a choice of "target" by pressing the right lever or "distractor" by pressing the left lever. Fixation was maintained for the duration of the observation period.

AP = 19, L = 22). By swivelling the guide tube different sites could be accessed within an approximately 10x10mm cortical region. Action potentials were amplified (Bak Electronics, Model 1A-B), and routed to an audio-monitor (Grass AM-8) and to a time-amplitude window discriminator (Bak Model DIS-1). The output of the window discriminator was used to trigger the real-time clock interface of a PDP11/83 computer.

2.4 Visual stimuli

The visual objects were presented on a monitor situated 97 cm from the animal. The selection of the vertices of the wire objects within a three-dimensional space was constrained to exclude intersection of the wire-segments and extremely sharp angles between successive segments, and to ensure that the difference in the moment of inertia between different wires remained within a limit of 10%. Once the vertices were selected the wire objects were generated by determining a set of rectangular facets covering the surface of a hypothetical tube of a given radius that joined successive vertices.

The spheroidal objects were created through the generation of a recursively-subdivided triangle mesh approximating a sphere. Protrusions were generated by randomly selecting a point on the sphere's surface and stretching it outward. Smoothness was accomplished by increasing the number of triangles forming the polyhedron that represents one protrusion. Spheroidal stimuli were characterized by the number, sign (negative sign corresponded to dimples), size, density and sigma of the gaussian type protrusions. Similarity was varied by changing these parameters as well as the overall size of the sphere.

Test-views were typically generated by ±10 to ±180 degree rotations around the vertical (Y), horizontal (X), or the two oblique ($\pm 45^o$) axes lying on the XY plane.

2.5 Data Analysis

Mean spike rates are distributed symmetrically, that is the mean is an accurate representation of central tendency coinciding with the median of the distribution. The significance of differences between mean spike rates measured during the target presentations and those measured during the distractor presentations can therefore be tested by using the non-parametric Walsh test for two related samples (Walsh, 1949). For our sample size (N = 9 presentations per target-view or distractor), the power-efficiency, *i.e.* approximately the percentage of the total available information per observation which is utilized by the test, of the one-tailed Walsh test at $\alpha = 0.011$ is 98% of that of the parametric t test at $\alpha = 0.05$, while it avoids the the use of assumption-laden dispersion measures. The neurons presented here as *view-selective* gave equal or greater responses to target views than to the views of the distractors, at $\alpha = 0.011(min[d3, \frac{1}{2}(d1 + d5)] > 0)$.

3 Results

3.1 View selectivity

Figure 2 describes the sequence of events that composes a single observation period. An observation period began with the presentation of a small fixation spot. Successful fixation was followed by the *learning phase*, during which the target was presented for 2 to 4 seconds from one viewpoint. This view of the target, called the *training view*, was presented in oscillatory motion $\pm 15^o$ around a fixed axis at 0.67Hz to provide the subject with complete $3D$ structure information. The *learning phase* was followed by a short fixation period after which the *testing phase* started. A *testing phase* consisted of up to 10 sequential trials, in each of which the test stimulus, a static view of either the target or a distractor, was presented. Thirty target views 12^o apart and 60 to 120 distractors were tested in a given session. The duration of stimulus presentation was 500-800 msec, and the monkeys were given 1500 msec to respond by pressing one of two levers: the right lever upon presentation of a target view and the left upon presentation of a distractor. Typical reaction times were below 1000 msec for both animals. An experimental session consisted of a sequence of 60 observation periods, each lasting about 25 seconds.

A total of 970 IT cells were recorded from two monkeys during combined psychophysical and electrophysiological experiments, in which the subject performed either a fixation task, or the recognition task described above. All data barring those shown in the last figure were collected using objects that the monkeys could recognize from any viewpoint (hit rate above 95% for all views, and false alarm below 5% for all distractors). The animals' view-invariant performance in the case of these objects was a result of training on multiple views, which lead to generalization around an entire axis, and eventually giving feedback for all views. A large majority of the isolated neurons were visually active when plotted with a variety of simple or complex stimuli, including some of the wire or spheroidal objects. Other neurons were inhibited by the presentation of target objects, and a small fraction of cells were inhibited by any stimulus including the fixation spot.

A number of units, however, responded selectively to a subset of views of one of the known target objects, firing much less or not at all for the distractors. The response of these neurons for different views was approximated by fitting to the data a gaussian function centered on the view eliciting the greatest response. If a cell responded to two subsets of views, as was the case for several cells, the linear sum of two gaussian functions, one centered on each "most effective" view, was used to fit the response. The standard deviation of these functions, which can be viewed as a measure of the generalization field of the cell, was used to classify the neurons based on the following criterion. Cells ($N = 61$) were considered selective if they responded significantly more to target views within two standard deviations of the preferred view, than for any of the distractors (see methods).

An example of a view-selective neuron is shown in Figure 3a. The cell's firing rate reached a maximum upon presentation of one particular object view and declined as the object was rotated away from this *preferred* view. Figure 3b shows sixteen out of the 60 tested distractor wire-objects and an associated histogram of the response each elicited. The within-class recognition task the animal was performing during the electrophysiological experiments provided an internal control against common or trivial features being responsible for the behavior of the neurons. Examination of the views of the target for which the cell is selective reveals a couple features that may be characteristic for that view of the target. For example, the inverted "V" (circled) in the 0^o view in Figure 3a, appears to be a prominent feature that all the response-eliciting target views have in common. Could the neuron simply be selectivly firing for the presence of this particular feature? This is not likely to be the case as an inverted "V" is also present in several of the distractors (see the circled regions of distractors 18, 25, 44, 49, 50 in Figure 3b).

Similar results were obtained with the class of spheroidal objects (Figure 4). Here, too, the neuron responds maximally to one view of the object, 72^o away from the zero-view, with its response declining as the angle of rotation deviates in either direction from the preferred view. Figure 4b shows the "best-response" eliciting distractors. Although all views of the target have one particular protrusion which remains visible in all views, this alone does not seem to be sufficient to elicit any sort of response. As indicated by the circled region of view "72^o", all of the views eliciting a significant response share the presence of a "face-like" region containing two dimples and a small protrusion in the lower right. However, similar regions are also present in two of the distractors, 12 and 14 in the bottom half of the figure, and neither of these elicit any activity from the cell whatsoever.

The generalization field of a number of view-selective neurons was examined for all rotations in depth using views neighboring the preferred view along all four axes. An example is shown in Figure 5a. This cell responded best to the 0^o view of the object and its response magnitude decreased with increasing angle of rotation along all axes. A small percentage of the view-selective cells (5 out of 61) exhibited their maximum discharge rate for two views 180 degrees apart (Figure 5b). The same pattern was observed in the behavioral performance of the monkeys for several objects (Logothetis et al., 1994). In both cases, this type of response was specific to wire-like objects whose zero and 180^o views appeared as mirror-symmetrical images of each other, due to accidental minimal self-occlusion.

Figure 6 shows the distribution of the generalization fields of view-selective cells for the wire-like and the spheroidal objects. The insets show the coefficients of determination indicating the goodness of fit. Both object-types gave similar tuning width, which was always less than or equal to the behavioral generalization field of monkeys trained with one view of similar objects (Logothetis et al., 1994).

A number of the objects used extensively during the training of the animal were also used during the electrophysiology sessions. For several of these objects, multiple neurons were found that were selective for different views of the same object. Figure 7a through 7d illustrates such a case for four units. Three out of the 970 cells responded selectively to specific objects presented from any viewpoint. Figure 7e shows such a neuron that appears to have properties of object-center descriptions. The cell responds about equally well for all target views and significantly less to any of the 120 distractors.

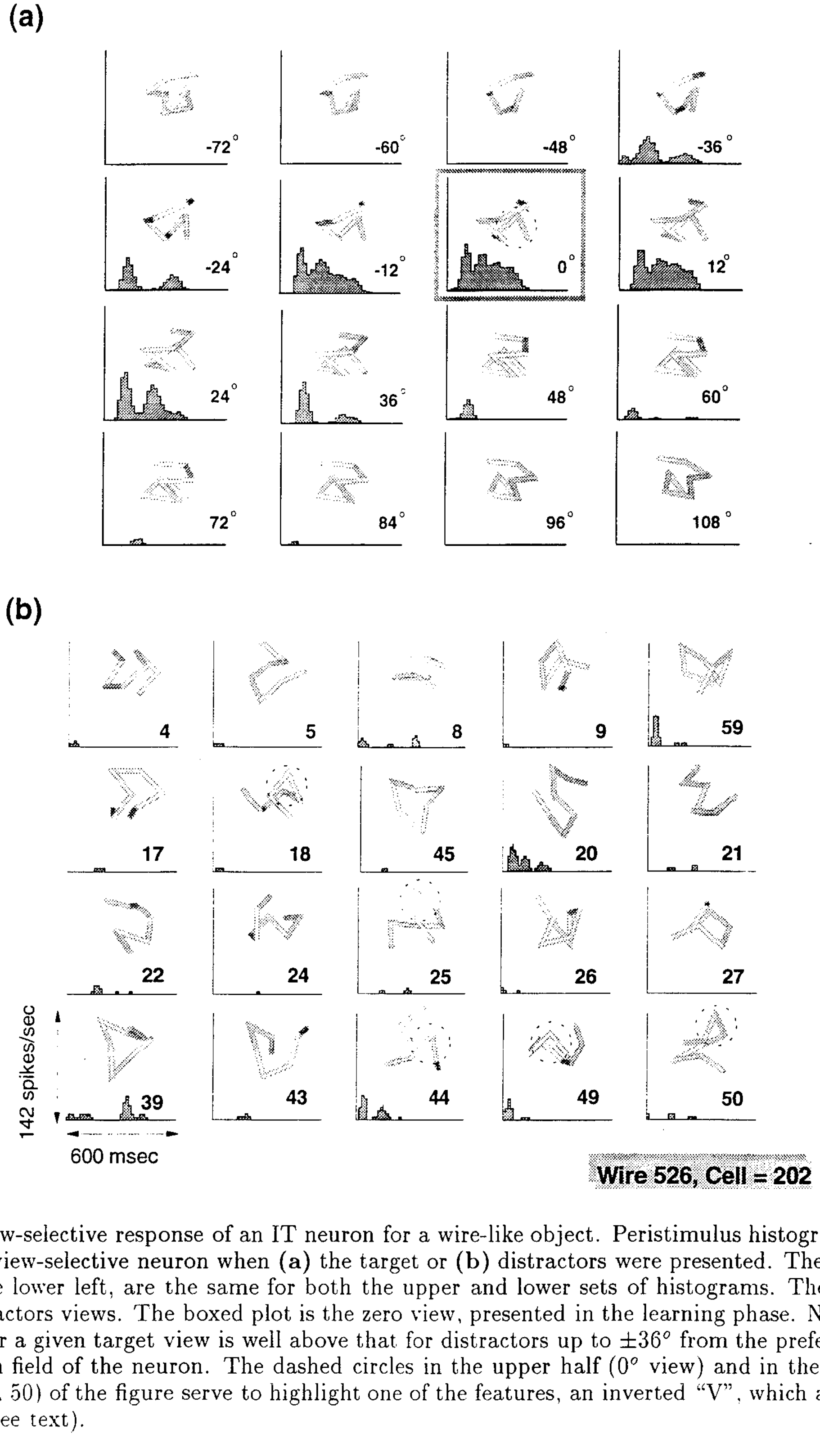

Figure 3: View-selective response of an IT neuron for a wire-like object. Peristimulus histograms (PSTHs) show the activity of a view-selective neuron when (a) the target or (b) distractors were presented. The ordinate and abscissa, labeled in the lower left, are the same for both the upper and lower sets of histograms. The insets show he target and the distractors views. The boxed plot is the zero view, presented in the learning phase. Note that the activity of the neuron for a given target view is well above that for distractors up to ±36° from the preferred view, defining the generalization field of the neuron. The dashed circles in the upper half (0° view) and in the lower half (distractors 18, 25, 44, 49, 50) of the figure serve to highlight one of the features, an inverted "V", which all of these images have in common (see text).

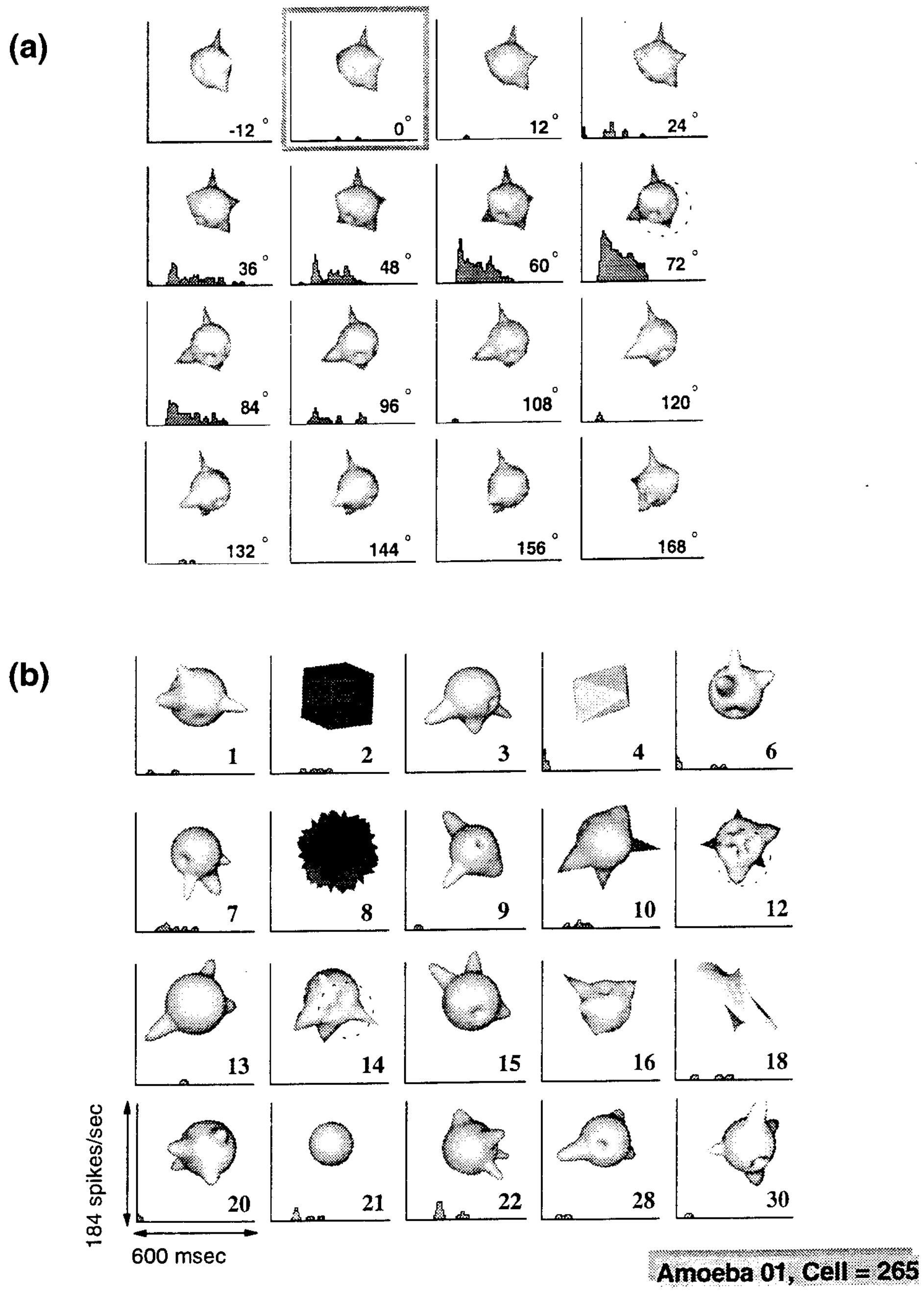

Figure 4: View-selective response of a neuron for a spheroidal object. Conventions as in Figure 3.

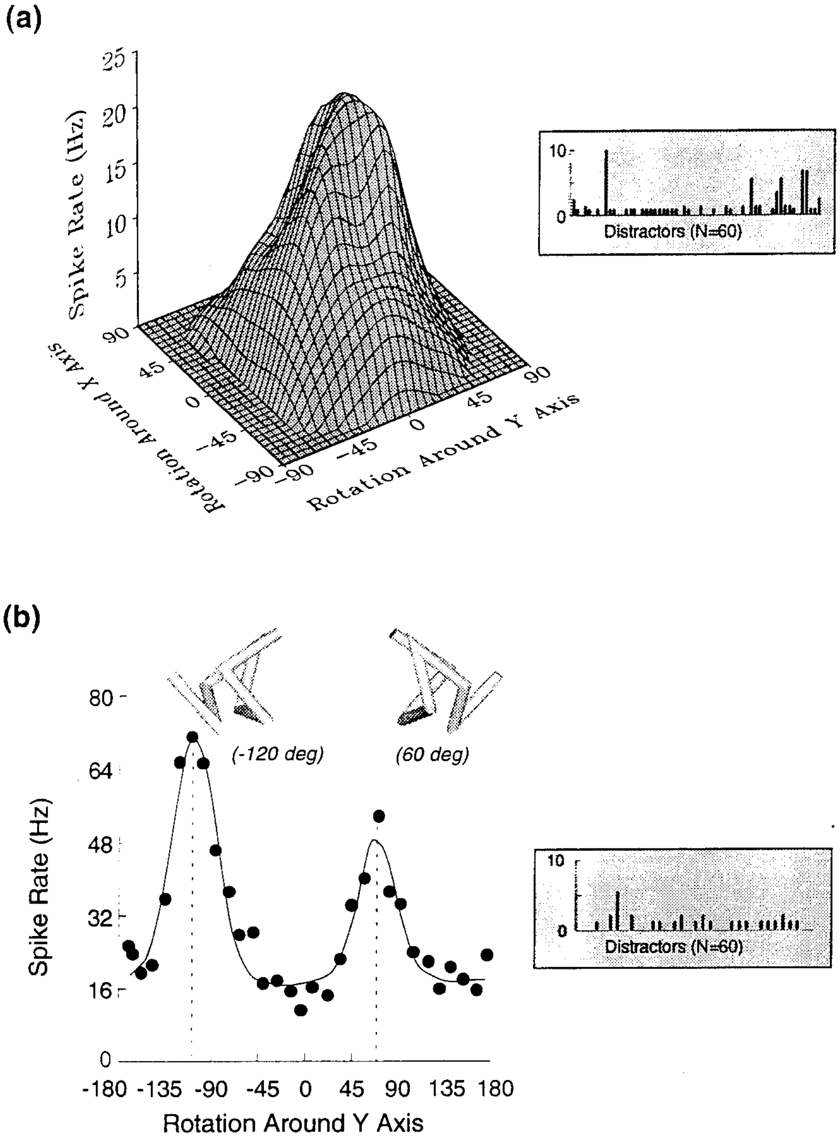

Figure 5: (a) Response of a view-selective neuron to rotations around the preferred view along four axes. The z-dimension of the plot is spike rate and the x and y dimensions show the degrees of rotation of the target object along either or both of these axes. The volume was generated by testing the cell's response for rotations out to ±60^o around the x and y axes as well as along the two diagonals. The magnitude of response fell of about the same for rotations away from 0^o along all of the axes tested. The activity of the neuron for the 60 distractors is shown in the inset. (b) Response of a neuron selective for pseudo-mirror-symmetric views, 180^o apart, of a wire-like object. The filled circles are the mean spike rates for target views around one axis of rotation. The solid black line is a DWLS-smoothed view-tuning curve. The two inset images depict the -120^o and 60^o views around both of which the neuron showed view-selective tuning. The activity of the neuron for the 60 different distractor objects used during testing is shown in the inset gray box.

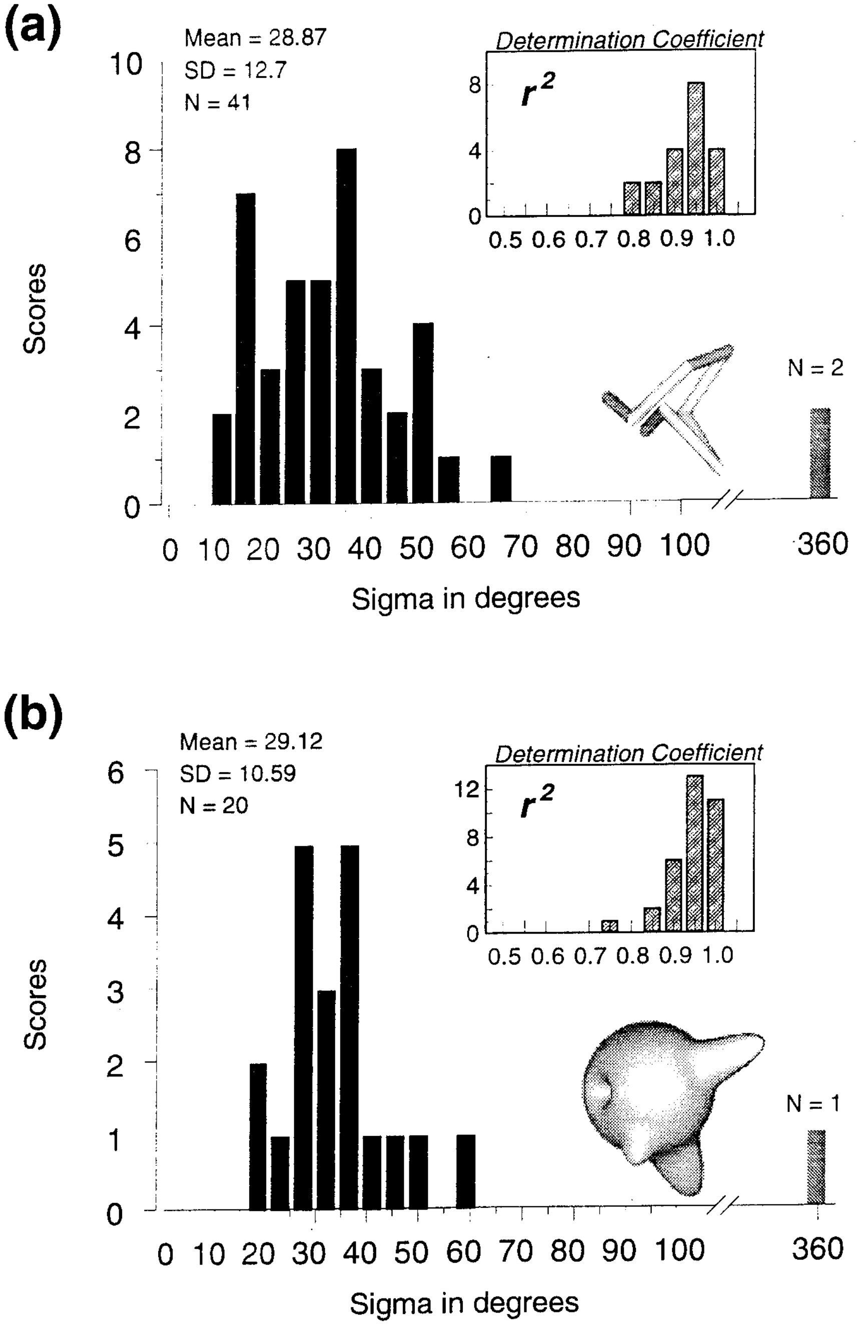

Figure 6: Distribution of the standard deviation of the gaussians fitted to the view-tuning curves of IT neurons for the wire-like (**a**) and the amoeba (**b**) objects. The black bars in both plots represent the 61 view-selective neurons. The gray bars show the three units that responded in a view-invariant manner for a given object. The insets show the coefficients of determination, indicating the goodness of the fit.

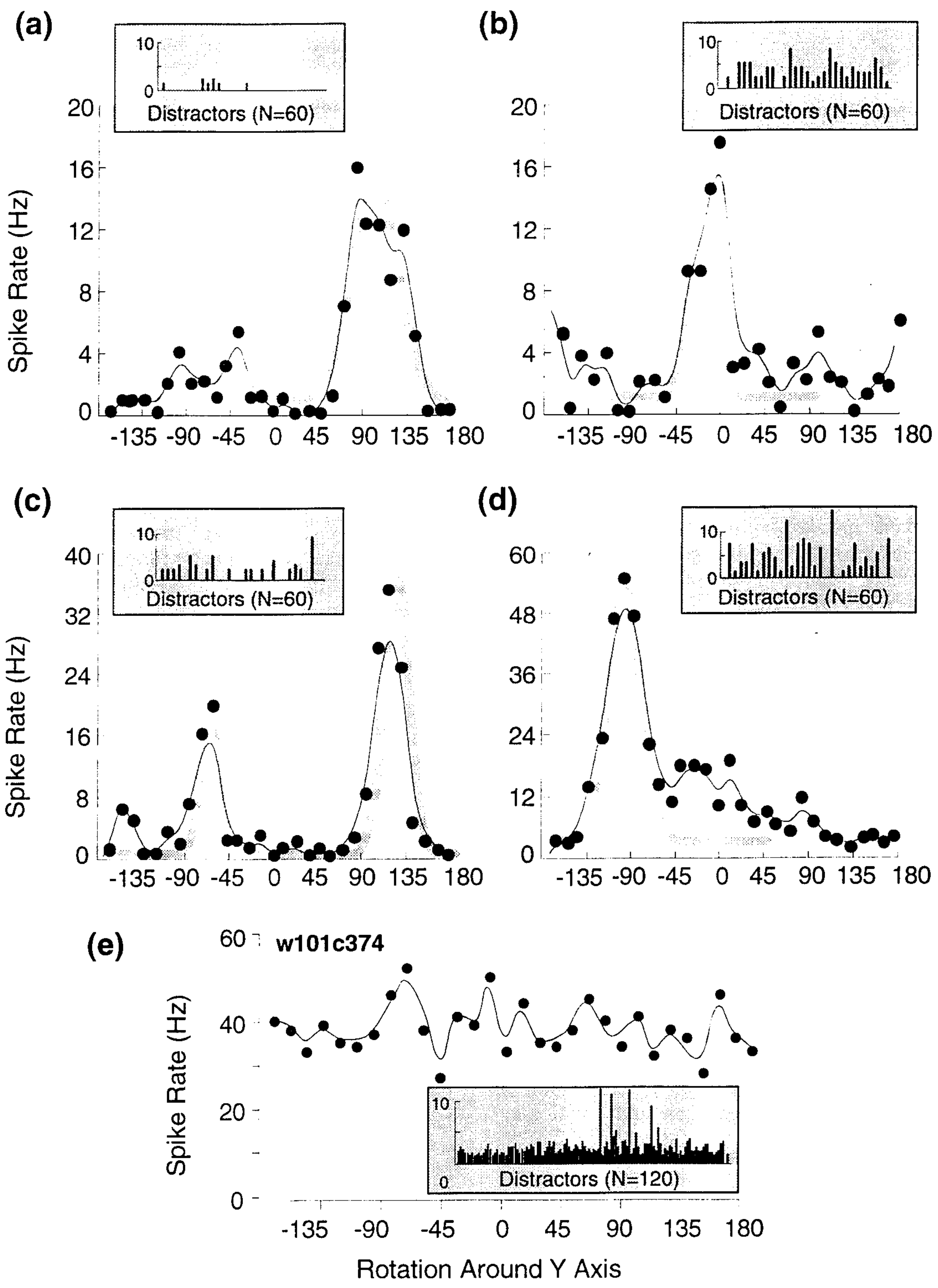

Figure 7: (a) - (d) View-selective responses of neurons tuned to different views of the same wire-object. All data come from the same animal (S5396). The filled circles are the mean spike rates (N=10), and the thin black lines DWLS-smoothed view-tuning curves. The thick gray lines are a nonlinear approximation of the data (QNMT) with the function $R(\theta) = \sum_{i=1}^{N} c_i exp(-(||\theta - \theta_i||)^2/2\sigma_i^2) + R_0$, where $N = 1$ or 2. (e) An example of a neuron showing view-invariant repsonse for a known wire object. The behavioral performance of the monkey for this object was view-independent due to its having been used as a training object (see text). The insets in (a) through (e) show the activity of the neuron the 60 or 120 distractors used during testing.

3.2 Translation and scale invariance

Among the population of neurons examined, we could identify a number of units that showed a large degree of size invariance. Figure 8 is an example of a view selective neuron the response of which was found to be invariant to changes in size. Whether the stimulus substended one degree of visual angle or six degrees the magnitude of the cells response was the same. Note that the fixation spot, the only unchanging part of the stimulus, did not elicit a response from the cell during the first 500ms of the trial before the stimulus onset. Figure 9 shows the response of the same cell when tested for positional invariance. In this case the center of the stimulus was translated 7.5 degrees from the fixation spot. With the exception of the brief on-transient, the cell's activity does not deviate from the baseline for all tested positions. Thus, this cell, while scale invariant, appears to be position dependent for relatively large displacements. The responses shown in Figures 8 and 9 were collected during a simple fixation task.

The response of eight view-selective neurons were tested for scale and translation invariance in the context of the object recognition task using the preferred view of the object. The stimulus sizes used subtended from 1.9 to 5.6 degrees of visual angle, and the positions were tested all at a radial distance of 3.15 degrees. An example of a view-selective neuron responding invariantly to changes in both size and position is shown in Figure 10.

This particular cell was selective when a limited region of the object around 120 degrees (Figure 10a) was presented, and responded 3.5 times more for the preferred target view than for the best distractor (Figure 10b). Responses to scaling and translation were tested using the preferred view. Figure 10c shows the ratio of the target response to the mean response for the ten best distractors for the sizes tested. Note that all of the distractors were of the default size and were presented foveally. The responses of the same cell to translation are plotted in Figure 10d. This particular neuron showed some variance in its response depending on stimulus position, however, in all cases its response for an eccentrically presented target was still at least twice that for foveally presented distractors. Seventy-five percent of the tested neurons gave only scale-invariant responses while 35% were invariant for both scale and position.

3.3 Responses to rotations in the view plane

Neurons were also tested for rotation in the view plane. Most units appeared to be orientation selective (Figure 11b). However, the initial performance of the animal also appeared to be orientation dependent for any given novel object rotated in the view plane (Figure 11a). In almost all cases, however, the initial generalization field for picture-plane rotations appears to be broader than that typically obtained for rotations in depth (Logothetis et al., 1994). Figure 11c illustrates the behavioral progression of one animal's recognition performance as it evolved from initially view-dependent to almost completely view-invariant for two different objects. Generalization performance often progressed rapidly, over the course of a few test sessions, to view-invariant performance. This is in strong contrast to the view-dependent performance seen for rotations in depth, which changed very little for the duration of testing (as many as fifteen sessions without feedback).

4 Discussion

The results of this study suggest an experience dependent plasticity in IT neurons, and support the idea of a population of neurons with configurational selectivity being a more general mechanism for encoding complex, "non-decomposable" objects. The neurons discussed above responded selectively to novel objects that the monkey had recently learned to recognize. None of these objects had any prior meaning for the animal, nor did they resemble anything familiar in the monkey's environment. View-selective responses were found for both object types tested and were not limited to any one single region of the an object. However, when cells were tested with objects, which the monkey could recognize only from a specific viewpoint, no selective responses were ever encountered for views that the animal systematically failed to recognize. The reported cell responses are unlikely to reflect a general sensation of familiarity or arousal, since the majority of the neurons responded selectively to a subset of the tested object-views, even when the animal's recognition performance was view-invariant (as in all cases except in Figure 11). Thus it seems that neurons in this area may develop complex, configurational selectivity as the animal is trained to recognize specific objects. Such neurons can be regarded as "blurred-templates", the tolerance of which to small rotations in depth represents a form of limited generalization. The capacity of some IT neurons to respond to both an object view and its "pseudo-mirror-symmetrical" view can be viewed as a broader form of generalization, possibly underlying the reflection-invariance observed during the psychophysical experiments (Logothetis et al., 1994). Distinguishing mirror images has no apparent usefulness to any animal, and the inability of normal children to distinguish between mirror-symmetrical letters or words (Orton, 1928; Corballis and McLaren, 1984) may be an adaptive mode of processing visual information, and not a "confusion" (Bornstein et al., 1978; Gross and Bornstein, 1978). In fact, theoretical and psychophysical work suggests that reflection-invariance facilitates the recognition of bilaterally symmetric visual objects (Vetter et al., 1994). Interestingly, neurons responding to mirror-images of a face appear very early in the visual system of the monkey (Rodman et al., 1993).

A significant number of neurons showed response invarinace to affine image transformations. Similar response behavior has been earlier reported for $2D$ patterns like the Fourier descriptors (Schwartz et al., 1983) and for faces (Desimone et al., 1984; Rolls and Baylis, 1986; Tovee et al., 1994). In our sample, position invariance varied from one extreme, where response was strongly reduced with small translation (often less than 2 degrees), to the other extreme where response remained largely invariant for eccentricites up to 7.5 degrees.

Surprising was the degree of view-dependency of the

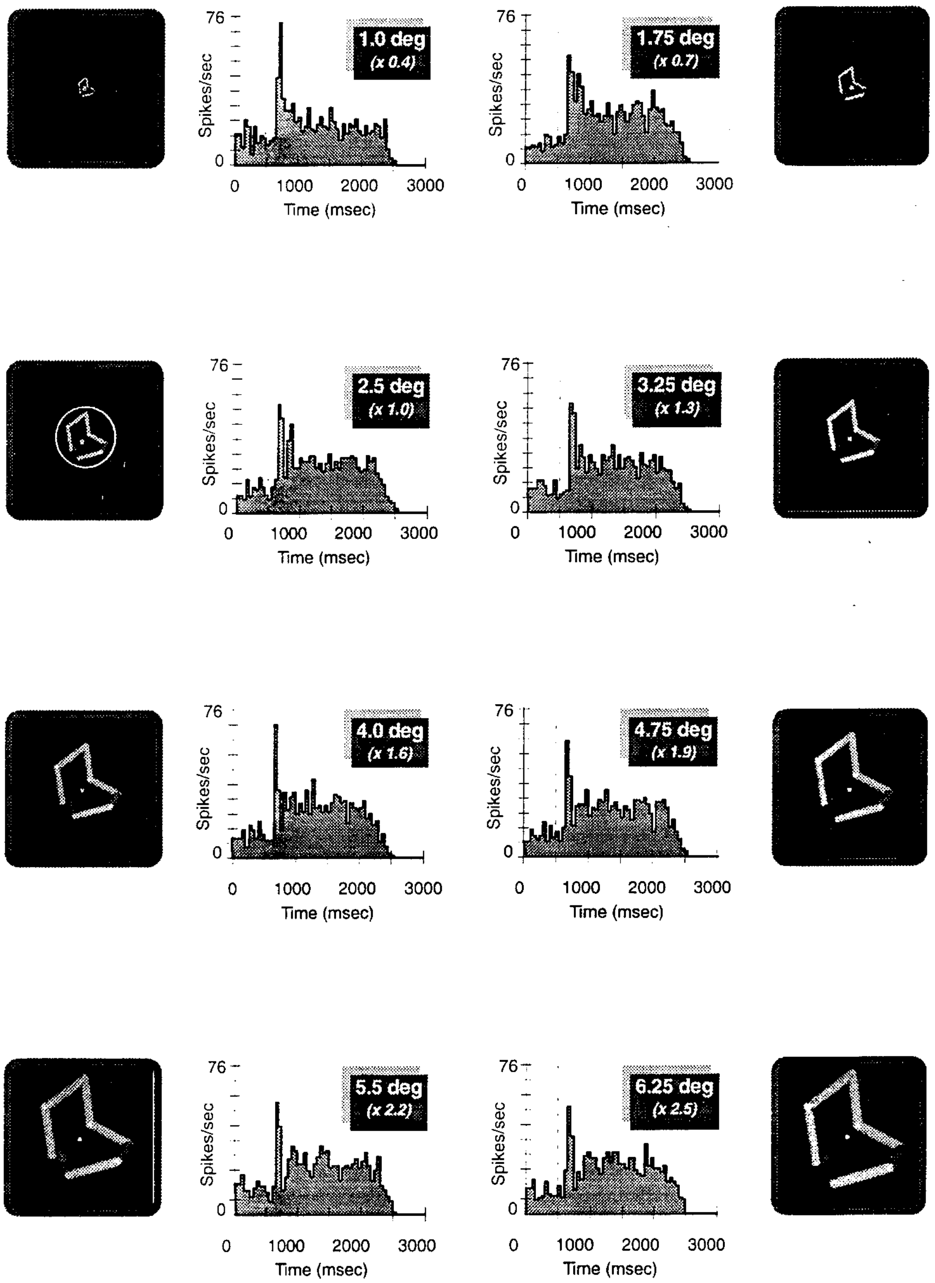

Figure 8: Response invariance to changes in size in a view-tuned neuron. The monkey was performing a simple fixation task in which each trial lasted 2500ms. PSTHs show the activity of the neuron over the course of a trial. The ordinate is spike rate and the abscissa is time. The animal fixated without a stimulus for the first 500ms at which point a stimulus would appear (indicated by the dashed line), and it continued to fixate for 2000ms, responding to a change in fixation spot color at the end of the trial. Each stimulus is shown to the side of its respective histogram. The circled stimulus is the one used for testing view-selectivity.

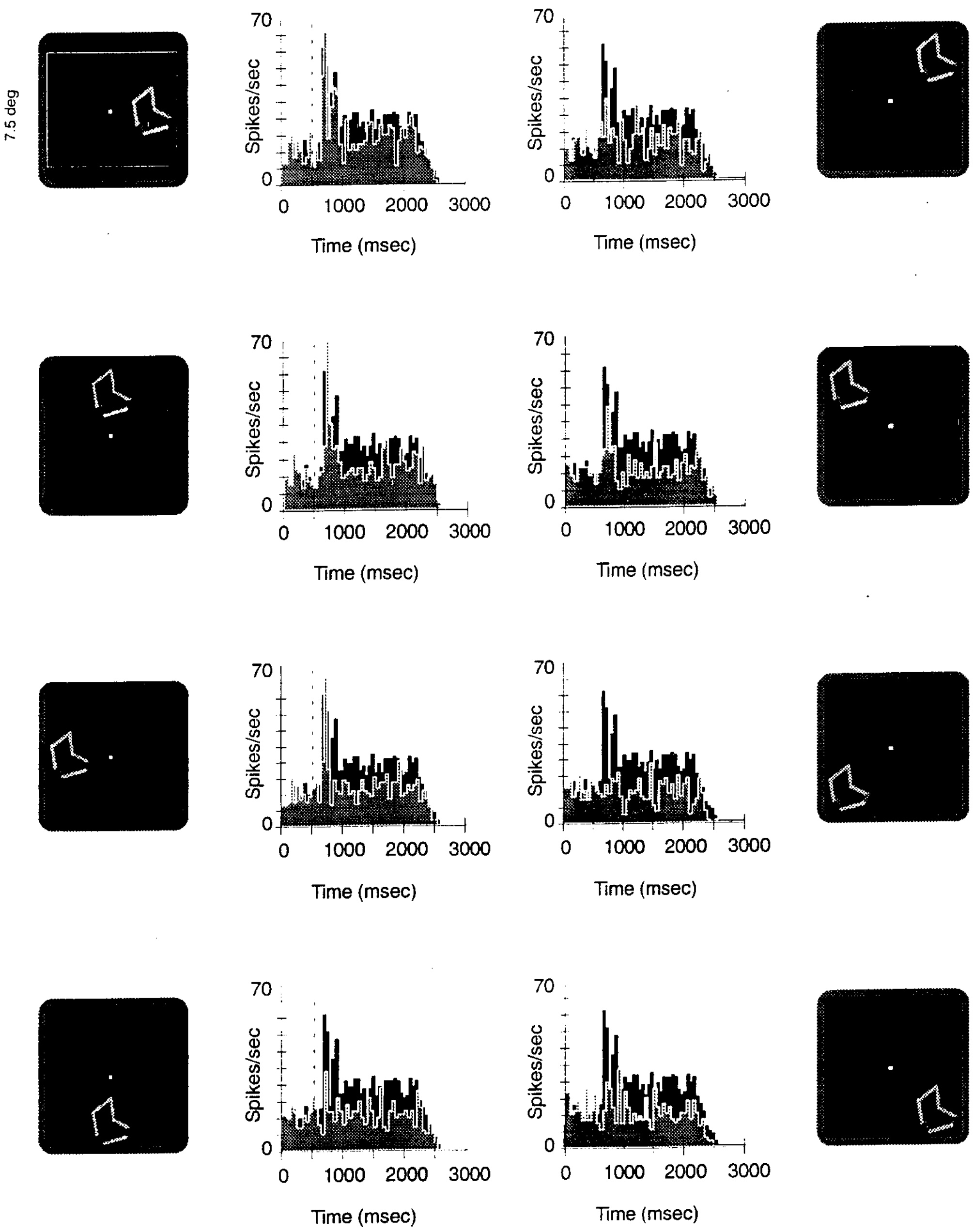

Figure 9: Responses to translation of an object in the picture-plane. Data are from the cell presented in Figure 7. The activity of the neuron for the default wire presented foveally (shown in Figure 7) is represented here by the black histogram in the background of each plot. The gray PSTHs show the activity of the cell for the eight positions tested. In each case the center of the wire was translated 7.5 degrees from the central fixation spot. Other than a short transient of activity, cell activity is barely distinguishable from baseline when the stimulus is presented at each of the eccentric positions. For smaller translations (less than 2 degrees), however, no such position dependence was observed.

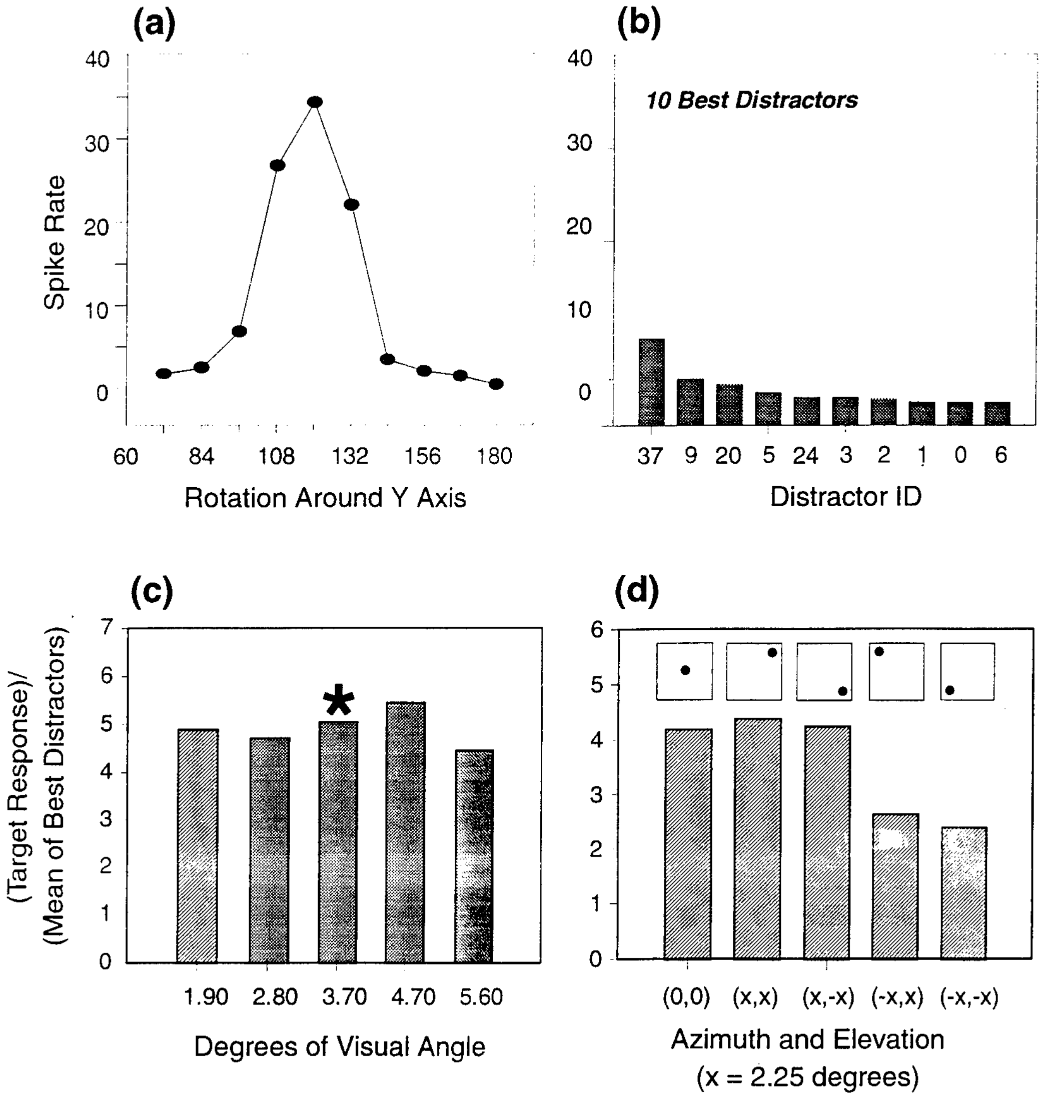

Figure 10: A view-selective neuron responding invariantly to changes in size size and position. (**a**) Tuning curve showing activity of the neuron for a limited region of the object. The preferred view corresponds to a 120° rotation of the object around the Y-axis. (**b**) The responses of the cell for the ten best distractors. Distractors were always presented foveally and at the default size. The best target view was used to examine the cell's response to changes in size (**c**) and position (**d**). The response of the cell is plotted in both graphs as a ratio of the mean-spike-rate for a target view to the mean of the mean-firing rates for the top ten distractors. The bar representing the response to the default size. is indicated by the asterisk in (**c**). The smallest size, 1.9°, was used to test translation. The ordinate of the graph indicates the position of each test image in terms of its azimuth and elevation.

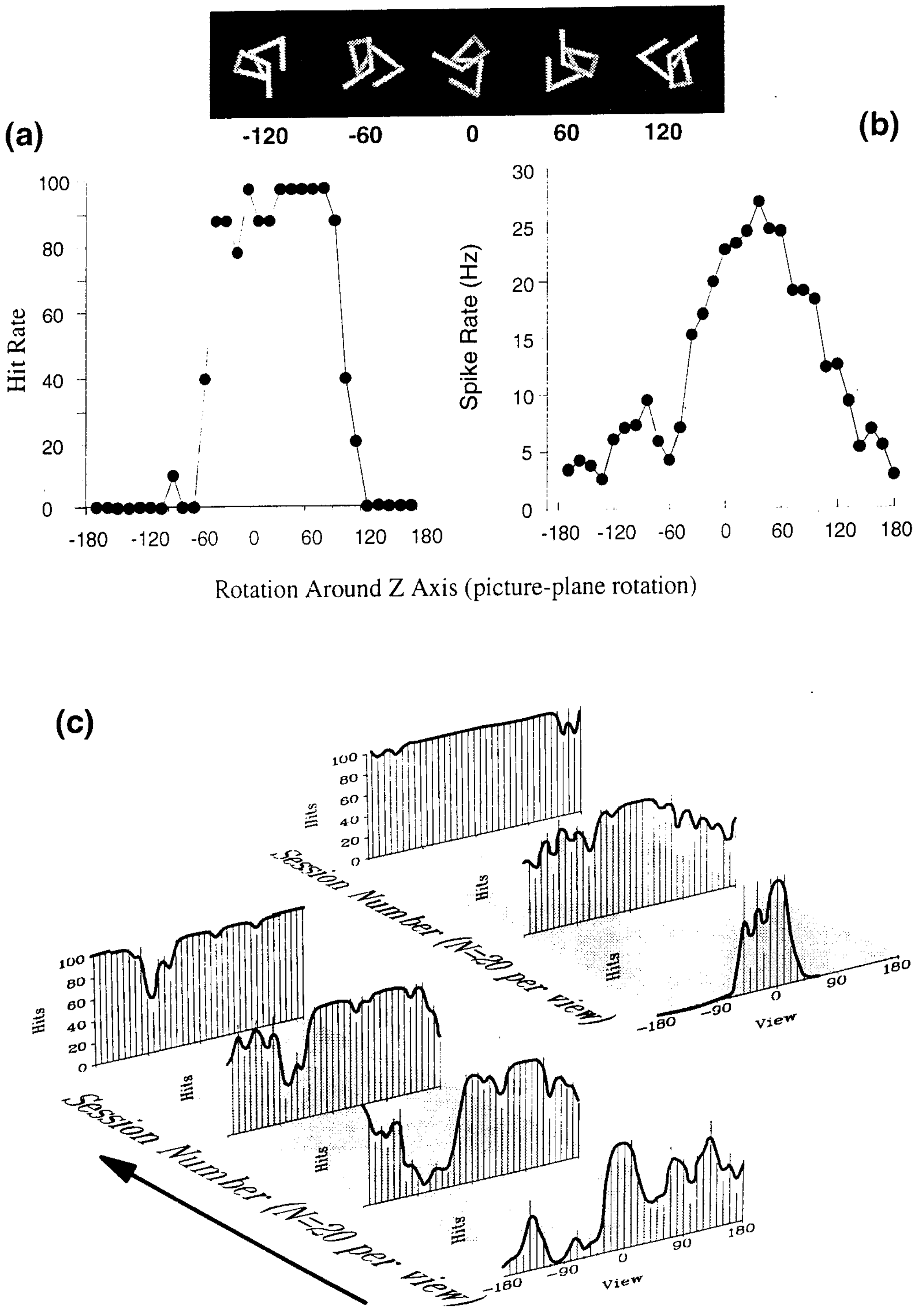

Figure 11: View-dependent behavioral performance and view-selective neuronal response for an image rotated in the picture-plane. (**a**) Performance of the animal in terms of hit rate (N = 9 trials per view). In this example, no training was given for the zero view prior to testing. (**b**) The plot depicts the view-tuning curve of the neuron in terms of mean-spike-rate. The abscissa of both plots is rotation angle. (**c**) Improvement of performance for recognition of views resulting from view-plane rotations. The X-axis is rotation angle, the Y-axis increasing session number, and the Z-axis hit rate. One test session included ten presentations of each target view, thirty-six in all, spaced at ten degree intervals. Each curve, starting in the front and proceeding to the back, illustrates the performance over two test session (N = 20 presentations of each target view). The animal was familiarized with the zero-view of the object during one brief training session prior to testing. No feedback was given during the testing periods as to the correctness of the response.

cell and the monkey responses for rotations in the plane of view. Psychophysical studies in humans have revealed that the recognition of objects rotated in the picture-plane is different than the recognition of objects rotated in depth. For example, Tarr and Pinker (Tarr & Pinker, 1989, 1990; Tarr and Pinker, 1991) studied the effects of rotation in the picture plane on recognition and found that familiarization with one view of an object results in view-independent performance, although reaction times do increase with deviation from the learned view. This performance can be altered by training the subjects briefly on a second view, resulting in an improvement in performance around the new learned view and to a lesser extent for those views between the two familiar views. In our experiments, the behavior of the monkeys was initially strongly view-dependent in terms of error rate. In contrast to the recognition performance observed for rotations of the object in depth, however, hit rate for view-plane rotations increased gradually over successive sessions without any feedback to the animal as to the correctness of its response. No neuron was isolated long enough to observe any possible changes at the single-cell level.

A question that arises from these results is: are such neurons really responding to the "views" of the tested objects? Studies by Tanaka and his colleagues (Tanaka et al., 1991) showed, for instance, that the response of many neurons to complex objects can be mimicked using simpler forms representing regions of the objects. In a similar vein, the neurons studied here could be responding to a reduced set of features of the wire or spheroidal objects and not to an entire view. Two observations seem to refute such an alternative. Firstly, the neurons were tested with a variety of simple objects, including geometric patterns of different orientations, that failed to elicit any response. Second, the presentation of between 60 and 120 distractors from the same or a different object class served as a selectivity-control for each of the targets. Thus in the case of the wire-objects, for example, given the largerly invariant responses of IT neurons for small translations (Tovee et al., 1994), the distractors had at least 60 different combinations of simple features like orientations, angles, or terminations, some of which were highly similar to those comprising the target object. As a matter of fact, several cells did respond to the presentation of the target and to a number of distractor objects, presumably excited by such simpler features. However, the selective cells discussed here gave minimal and sometimes no response for distractor objects, even when the latter shared a few characteristic regions with the target, indicating that a specific organization of some features was required for eliciting the neuron's response.

Nevertheless, both arguments are based on qualitative observations, and what we present here as "view-selectivity" may still be reducible to less complex feature constellations. A systematic, mathematical analysis of object-views that elicit similar neural responses, and an attempt to develop algorithms for biologically-plausible image decomposition may provide an answer to the *selectivity* question, and this is the focus of current experiments.

5 Conclusions

Taken together, these data suggest the possibility of a recognition architecture similiar to that schematically described in Figure 1. The discharge rate of many IT neurons was found to be a bell-shaped function of orientation centered on a preferred view. A very small number of neurons exhibited object-specific but view-invariant responses that might be the result of the convergence of view-dependent units into neurons showing characteristics of object-centered descriptions. The input of each view-selective unit can be considered as the conjunction of simpler features extracted at earlier stages in the visual system. The variability in the degree of response invariance during affine image transformations also hints to a multilayer, possibly hierachical architecture.

Such a scheme is obviously oversimplified and lacks top-down mechanisms that strongly affect recognition performance. The processing of object information is undoubtedly far more complex, and representations might be local and explicit or distributed and implicit according to the recognition task or the stimulus context. Although the ultimate goal of a recognition system is to describe grouped object-features in a more abstract format that captures the invariant, three-dimensional, geometric properties of an object, early representations may be in some cases strongly configurational. Moreover, for visually complex, non-decomposable objects, like many biologically meaningful objects, holistic representations may be the only ones possible. Neurons selective for object-views and tolerant of varying extents of image transformations may then be elements of one possible mechanism for such representations.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychol Rev, 94*, 115–147.

Bornstein, M., Gross, C., & Wolf, J. (1978). Perceptual similarity of mirror images in infancy. *Cognition, 6*, 89–116.

Bruce, C., Desimone, R., & Gross, C. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol, 46*, 369–384.

Brunelli, R., & Poggio, T. (1991a). Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*, 1042–1052.

Brunelli, R., & Poggio, T. (1991b). HyberBF Networks for Real Object Recognition. In J. Mylopoulos, & R. Reiter (Eds.), *Proc. 12th Intl. Joint Conf. on Artificial Intelligence (IJCAI)* (pp. 1278–1284). Sydney, Australia: Morgan Kaufman.

Bülthoff, H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc Natl Acad Sci U S A, 89*, 60–64.

Corballis, M., & McLaren, R. (1984). Winding one's Ps and Qs: Mental rotation and mirror-image discrimination. *J Exp Psychol [Hum Percept]*, *10*, 318–327.

Damasio, A. (1990). Category-related recogntion defects as a clue to the neural substrates of knowledge. *Trends Neurosci*, *13*, 95–99.

Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci*, *4*, 2051–2062.

Edelman, S., & Bülthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Res*, *32*, 2385–2400.

Farah, M. (1985). Psychophysical Evidence for a Shared Representation Medium for Mental Images and Percepts. *J Exp Psychol [General]*, *114*, 91–103.

Farah, M., McMullen, P., & Meyer, M. (1991). Can Recognition of Living Things be Selectively Impaired. *Neuropsychologica*, *29*, 185–193.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, *360*, 343–346.

Gross, C., & Bornstein, M. (1978). Left and Right in Science and Art. *Leonardo*, *11*, 29–38.

Gross, C., Roche-Miranda, C., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J Neurophysiol*, *35*, 96–111.

Jolicoeur, P., Gluck, M., & Kosslyn, S. (1984). Pictures and Names: Making the Connection. *Cogn Psychol*, *16*, 243–275.

Logothetis, N., Pauls, J., Bülthoff, H., & Poggio, T. (1994). View-dependent Object Recognition in the Primate. *Curr Biology*, *4*, 401–414.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman & Company.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–820.

Orton, S. (1928). Specific reading disability - strephosymbolia. *JAMA*, *90*, 1095–1099.

Perrett, D. (1985). Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. *Behav Brain Res*, *16*, 153–170.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hietanen, J., & Ortega, J. (1989). Frameworks of Analysis for the Neural Representation of Animate Objects and Actions. *J Exp Biol*, *146*, 87–113.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*, 263–266.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.

Richmond, B., Optican, L., Podell, M., & Spitzer, H. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. I. Response characteristics. *J Neurophysiol*, *57*, 132–146.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cogn Psychol*, *19*, 280–293.

Rock, I., DiVita, J., & Barbeito, R. (1981). The effect on form perception of change of orientation in the third dimension. *J Exp Psychol*, *7*, 719–732.

Rodman, H., Scalaidhe, S., & Gross, C. (1993). Response properties of neurons in temporal cortical visual areas of infant monkeys. *J Neurophysiol*, *70*, 1115–1136.

Rolls, E. (1984). Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum Neurobiol*, *3*, 209–222.

Rolls, E., & Baylis, G. (1986). Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp Brain Res*, *65*, 38–48.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cogn Psychol*, *8*, 382–439.

Schwartz, E., Desimone, R., Albright, T., & Gross, C. (1983). Shape recognition and inferior temporal neurons. *Proc Natl Acad Sci U S A*, *80*, 5776–5778.

Tanaka, J., & Taylor, M. (1991). Object Categories and Expertise: Is the Basic Level in the Eye of Beholder?. *Cogn Psychol*, *23*, 457–482.

Tanaka, K., Saito, H.-A., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol*, *66*, 170–189.

Tarr, M., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cogn Psychol*, *21*, 233–282.

Tarr, M., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame?. *Psychol Sci*, *1*, 253–256.

Tarr, M., & Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: Further issues. *Psychol Sci*, *2*, 207–209.

Tovee, M., Rolls, E., & Azzopardi, P. (1994). Translation Invariance in the Responses to Faces of Single Neurons in the Temporal Visual Cortical Areas of the Alert Macaque. *J Neurophysiol*, *72*, 1049–1061.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193–254.

Vetter, T., Poggio, T., & Bülthoff, H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Curr Biol*, *4*, 18–23.

Wachsmuth, E., Oram, M., & Perrett, D. (1994). Recognition of Objects and Their Component Parts: Responses of Single Units in the Temporal Cortex of Macaque. *Cereb Cortex*, *5*, 509–522.

Walsh, J. (1949). Some Significance Tests for the Median which are Valid under very General Conditions. *J Amer Statist Ass*, *44*, 64–81.

Yamane, S., Kaji, S., & Kawano, K. (1988). What facial features activate face neurons in the inferotemporal cortex of the monkey?. *Exp Brain Res*, *73*, 209–214.

REPORT DOCUMENTATION PAGE

Form Approved
OBM No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)

2. REPORT DATE
March 1995

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Spatial Reference Frames for Object Recognition Tuning for Rotations in Depth

5. FUNDING NUMBERS
N00014-93-1-0209,
NIH1R01EY10089-01,
N00014-93-1-0385,
NSF ASC-92-17041

6. AUTHOR(S)
Nikos K. Logothetis, Jon Pauls, and Tomaso Poggio

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Square
Cambridge, Massachusetts 02139

8. PERFORMING ORGANIZATION REPORT NUMBER
AIM 1533
CBCL 120

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Information Systems
Arlington, Virginia 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
None

12a. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

The inferior temporal cortex (IT) of monkeys is thought to play an essential role in visual object recognition. Inferotemporal neurons are known to respond to complex visual stimuli, including patterns like faces, hands, or other body parts. What is the role of such neurons in object recognition? The present study examines this question in combined psychophysical and electrophysiological experiments, in which monkeys learned to classify and recognize novel visual 3D objects. A population of neurons in IT were found to respond selectively to such objects that the monkeys had recently learned to recognize. A large majority of these cells discharged maximally for one view of the object, while their response fell off gradually as the object was rotated away from the neuron's preferred view. Most neurons exhibited orientation-dependent responses also during view-plane rotations. Some neurons were found tuned around two views of the same object, while a very small number of cells responded in a view-invariant manner. For five different objects that were extensively used during the training of the animals, and for which behavioral performance became view-independent, multiple cells were found that were tuned around different views of the same object. No selective responses were ever encountered for views that the animal systematically failed to recognize. The results of our experiments suggest that neurons in this area can develop a complex receptive field organization as a consequence of extensive training in the discrimination and recognition of objects. Simple geometric features did not appear to account for the neurons' selective responses. These findings support the idea that a population of neurons -- each tuned to a different object aspect, and each showing a certain degree of invariance to image transformations -- may, as an assembly, encode complex 3D objects. In such a system, several neurons may be active for any given vantage point, with a single unit acting like a blurred template for a limited neighborhood of a single view.

14. SUBJECT TERMS
AI, MIT, Artificial Intelligence, Monkey, Primate, Recognition, Learning, Memory, Inferotemporal Cortex, Cortical Physiology

15. NUMBER OF PAGES
18

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED